



ARL-MR-0891 • JUN 2015



# Wavelet Analysis for Molecular Dynamics

by Berend Christopher Rinderspacher, Jaydeep P Bardhan,  
and Ahmed E Ismail

## **NOTICES**

### **Disclaimers**

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.



# Wavelet Analysis for Molecular Dynamics

**by Berend Christopher Rinderspacher**

*Weapons and Materials Research Directorate, ARL*

**Jaydeep P Bardhan**

*Dept. of Mechanical and Industrial Engineering, Northeastern University*

*Boston, MA*

**Ahmed E Ismail**

*Dept. of Mechanical Engineering, RWTH Aachen University*

*Aachen, Germany*

<b>REPORT DOCUMENTATION PAGE</b>				<b>Form Approved OMB No. 0704-0188</b>
<p>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p><b>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</b></p>				
<b>1. REPORT DATE (DD-MM-YYYY)</b> June 2015	<b>2. REPORT TYPE</b> Progress	<b>3. DATES COVERED (From - To)</b> October 2013–September 2014		
<b>4. TITLE AND SUBTITLE</b> Wavelet Analysis for Molecular Dynamics		<b>5a. CONTRACT NUMBER</b> <b>5b. GRANT NUMBER</b> <b>5c. PROGRAM ELEMENT NUMBER</b>		
<b>6. AUTHOR(S)</b> Berend Christopher Rinderspacher, Jaydeep P Bardhan, and Ahmed E Ismail		<b>5d. PROJECT NUMBER</b> <b>5e. TASK NUMBER</b> <b>5f. WORK UNIT NUMBER</b>		
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b> US Army Research Laboratory ATTN: RDRL-WMM-G Aberdeen Proving Ground, MD 21005-5066		<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b> ARL-MR-0891		
<b>9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b>		<b>10. SPONSOR/MONITOR'S ACRONYM(S)</b> <b>11. SPONSOR/MONITOR'S REPORT NUMBER(S)</b>		
<b>12. DISTRIBUTION/AVAILABILITY STATEMENT</b> Approved for public release; distribution is unlimited.				
<b>13. SUPPLEMENTARY NOTES</b> primary author's email: <berend.c.rinderspacher.civ@mail.mil>.				
<b>14. ABSTRACT</b> We present a multiresolution diffusion-wavelet approach to operator compression, which does not require explicit preparation of atomistic-to-coarse-grained mappings. Our method takes as input the topology and sparsity of the bonding structure of a molecular system, and returns a hierarchical set of system-specific coarse-grained (CG) variables. The hierarchical compression provides a clear framework for modeling at many model scales (levels), beyond the common 2-level CG representation. Our results show that the hierarchy separates localized modes from larger-scale motions, e.g., long-range concerted backbone vibrational modes. Our approach correctly captures small-scale chemical features as well as large-scale features of the backbone. In particular, the new method's finest-scale modes are comparable to united atom models and other chemically defined CG models. Modes at coarser levels describe increasingly large connected portions of the target polymers. For polyethylene and polystyrene, spatial coordinates and their associated forces were compressed by up to 2 orders of magnitude. The compression in forces is of particular interest as this allows larger time steps as well as reducing the number of degrees of freedom.				
<b>15. SUBJECT TERMS</b> multiscale, multiresolution, wavelets, molecular dynamics, materials science				
<b>16. SECURITY CLASSIFICATION OF:</b>		<b>17. LIMITATION OF ABSTRACT</b>	<b>18. NUMBER OF PAGES</b>	<b>19a. NAME OF RESPONSIBLE PERSON</b> B Christopher Rinderspacher
<b>a. REPORT</b> Unclassified	<b>b. ABSTRACT</b> Unclassified	<b>c. THIS PAGE</b> Unclassified	UU	28
<b>19b. TELEPHONE NUMBER (Include area code)</b> 410-306-2811				

## Contents

---

<b>List of Figures</b>	<b>iv</b>
<b>1. Introduction</b>	<b>1</b>
<b>2. Methodology</b>	<b>2</b>
2.1 Basic Multiresolution	4
<b>3. Results</b>	<b>6</b>
3.1 Adaptive Multiresolution	8
3.1.1 Reconstruction Theory	8
3.1.2 A Priori Approximations for Reconstruction	10
3.2 Simulation Results and Analysis	12
<b>4. Concluding Remarks</b>	<b>17</b>
<b>5. References</b>	<b>18</b>
<b>List of Symbols, Abbreviations, and Acronyms</b>	<b>20</b>
<b>Distribution List</b>	<b>21</b>

## List of Figures

---

Fig. 1 Heat map of selected wavelet degrees of freedom (DoFs) in 1,4-D-glucose tetramer. Blue denotes positive coefficients, red negative. From left to right: fine-grained wavelet DoF covering a single repeat unit; coarse-grained wavelet DoF covering 1 half of the oligomer; coarser-wavelet DoF covering the oligomer isolating repeat units with sign changes; coarsest wavelet. .... 12

Fig. 2 Selected reduced wavelet representations of a polyethylene crystal. From a) to d) wavelet information is successively added. a) Coarse representation highlights the anisotropy in 1-dimensional chain averaged over all chain segments; b) coarse representation of the 2-dimensional chain plane; c) 3-dimensional representation; and d) full resolution. .... 13

Fig. 3 Fourier transform of the y-component of 1,000 atoms in crystalline polyethylene (100,800 atoms) excluding zero frequency to allow detail at other frequencies. Molecular dynamics at 500 K and 1 atm. Left: Individual power spectra per atom. Right: Power spectrum of magnitude of optimal representation. .... 13

Fig. 4 Fourier transform of the y-component of a 100,800-atom crystalline polyethylene system sampled at 1 fs. 3 scales are shown: 1st (left, finest scale), 5th (middle), and 12th scale (out of 25, right). .... 13

Fig. 5 Fourier transform of the z-component of alanine dipeptide in vacuum excluding zero frequency to allow detail at other frequencies. Molecular dynamics at 500 K and 1 atm. Left: Individual power spectra per atom. Right: Power spectrum of magnitude of optimal representation. .... 14

Fig. 6 Fourier transforms of alanine dipeptide degrees of freedom (DoFs) time series in vacuum excluding zero frequency to allow detail at other frequencies. Molecular dynamics at 500 K and 1 atm. Upper-left: z-component of optimal representation; Upper-right: 2nd finest optimal wavelet DoF z-component; Lower-left: 3rd finest optimal wavelet DoF z-component; Lower-right: 4th finest optimal wavelet DoF z-component. .... 15

Fig. 7 Top: Heat maps of the degree of freedom (DoF) correlations of a water dimer in vacuum; Bottom: Heat maps of the DoF correlations of alanine dipeptide in vacuum; Left: Correlations of the cartesian DoFs; Right: Correlations of the wavelet DoFs. .... 16

## 1. Introduction

---

Many problems in materials design involve optimizing semi-crystalline polymeric materials with respect to their dynamical behavior or nonequilibrium properties. These problems pose substantial modeling challenges that do not afflict crystalline or amorphous materials or raise questions about thermodynamics. First, length scales on the order of 10  $\mu\text{m}$  are required to model many semi-crystalline materials or rare-event phenomena involving low- to medium-defect concentrations (e.g., aggregation, crack initiation, and strain-induced crystallization). This scale exceeds current computational resources for molecular dynamics (MD) simulations by a factor of 1,000. Ultra-high-molecular-weight polyethylene offers a classic example of the scale challenge: despite its simple chemical makeup,  $\text{C}_n\text{H}_{2n+2}$  with  $n \gg 100,000$ , as a material it exhibits large regions of disorder that coexist with crystalline regions.

Compounding the challenge, fully atomic detail is required for performance-determining phenomena, such as vacancies, dislocations, and crack tips; consequently, MD simulations represent the state of the art for computational design of materials from first principles. In contrast, for crystalline materials one may use multiscale models that couple atomistic and continuum theories, e.g., the quasi-continuum model and the finite-element atomistic method.

We now highlight the main computational challenges faced by MD methods. First, high-frequency events, such as bond vibrations, dominate the state space sampled during an MD simulation (and the computational work performed in the process). Slower processes, such as torsions and translations, are usually the more relevant ones. The polymers' long chain lengths, and low defect concentrations, compound these effects by requiring a large number of degrees of freedom (DoFs). Similarly, the size of the crystalline or disordered regions is frequently too large for fully atomistic MD simulations, and hence MD is generally restricted to homogeneous systems, i.e., fully crystalline or wholly amorphous materials. A polymer melt with  $N$  particles requires  $O(N^3)$  time to equilibrate, making simulation essentially impossible without advanced simulation approaches.

Coarse-graining is an effective approach to reduce simulation cost. Coarse-grained (CG) simulations are usually faster than their atomistic counterparts due partially to the reduced DoFs and partially to an increased time step enabled by the elimination

of high-frequency vibrations. CG methods generally reduce predefined chemical groups, or sets of atoms, to a single “bead,” and the assignment of atoms and the bead properties are developed through chemical intuition and careful testing. The force-matching method of Lu et al.<sup>1</sup> and the reversible CG method for phenolic polymers by Harmandaris et al.<sup>2</sup> have greatly impacted accessible simulation scales for these materials. Other groups have used wavelet bases to construct potentials that are efficiently evaluated.<sup>3</sup> However, the human insights required at the atomistic level makes it difficult to develop higher-level CG models, which are essential to reach the desired 10- $\mu\text{m}$  length scale and commensurate time scales. Furthermore, these CG approaches introduce the difficult problem of generating consistent atomistic reconstructions, because particle-like coarse-grained beads lose information about the particles they subsume. Few models provide the on-the-fly adaptivity required for important problems, such as modeling crack initiation and propagation, or interfacial phenomena.

In the present work, we introduce a wavelet-based approach to extend the work of Ismail et al.,<sup>4,5</sup> which provides a consistent and systematic framework to derive multiple levels of model resolution while also reducing simulation complexity. Importantly for the dynamical and nonequilibrium metrics of interest, this approach captures molecular information relevant to kinetics in addition to thermodynamics. Our approach is tied strictly to the underlying physics and has the potential to increase accessible simulation sizes and durations by multiple orders of magnitude. Earlier wavelet approaches for MD addressed time-series analysis, a classical application of wavelet techniques, and did not explore model acceleration and approximation.

## 2. Methodology

---



---

The foundations of MD lie in the application of Newtonian mechanics to the energy functional

$$E = \frac{1}{2}M\|\dot{x}\|^2 + V(x), \quad (1)$$

where  $x$  are particle positions,  $M$  is the diagonal matrix of particle masses, and  $V$  is the applied potential. For the macromolecular systems we are interested in,  $V$  is usually partitioned into

$$V = V_{\text{bond}} + V_{\text{angle}} + V_{\text{torsion}} + V_{\text{non-bonded}};$$

$\mathbf{V}_{\text{bond}}$  is generally a pairwise harmonic potential:

$$V_{\text{bond}}(x) = \sum_{i \neq j} \frac{1}{2} k_{ij} (||x_i - x_j|| - r_0^{(ij)})^2.$$

The matrix  $K = (k_{ij})_{ij}$  is highly sparse. The atoms and  $K$  define a graph in which the atoms are the vertices and an edge between atom  $i$  and  $j$  has weight from  $K_{ij}$ . The maximum degree of a vertex in organic materials is 4, and even in organometallic complexes the coordination number is unlikely to exceed 6.

$\mathbf{V}_{\text{angle}}$  is a 3-body potential. Several functional forms are common for angle potentials complicating not only implementation but also choice of approximation. In all cases, the contributions vary much more slowly than in bonds, and the energies from angle interactions are typically an order of magnitude smaller than bond energies.

$\mathbf{V}_{\text{torsion}}$  is a 4-body potential that includes the dihedral and improper plane deformation interactions. Force constants tend to be 2 to 3 magnitudes lower than bond constants.

$\mathbf{V}_{\text{non-bonded}}$  includes all nonbonded interactions, such as electrostatics and van der Waals terms.

The matrix  $K$  is highly sparse. The atoms and  $K$  define a graph in which the atoms are the vertices and an edge between atom  $i$  and  $j$  has weight from  $K_{ij}$ . The maximum degree of a vertex in organic materials is 4, and even in organometallic complexes, the coordination number is unlikely to exceed 6. From Eq. 1, we derive equations of motion,

$$M^{\frac{1}{2}}\ddot{r} = M\ddot{x} = -\nabla V(x) = -M^{\frac{1}{2}}\tilde{\Delta}M^{\frac{1}{2}}x - M^{\frac{1}{2}}\tilde{V}(M^{\frac{1}{2}}x), \quad (2)$$

where  $\tilde{\Delta} := M^{-\frac{1}{2}}(\text{diag}(K\mathbf{1}) - K)M^{-\frac{1}{2}}$ ,  $\mathbf{1}$  is the all ones vector, and  $\text{diag}$  maps a vector to its corresponding diagonal matrix. By letting  $r = M^{\frac{1}{2}}x$ , the kinetic energy reduces to  $\frac{1}{2}||\dot{r}||^2$  and the equations of motion become

$$\ddot{r} = -\tilde{\Delta}r - \tilde{V}(r).$$

If  $\tilde{V} = 0$ , the system can be solved analytically,

$$\begin{pmatrix} r(t) \\ \dot{r}(t) \end{pmatrix} = e^{\begin{pmatrix} 0 & I \\ -\tilde{\Delta} & 0 \end{pmatrix} t} \begin{pmatrix} r_0 \\ \dot{r}_0 \end{pmatrix}.$$

Since  $\tilde{\Delta}$  is positive semi-definite, the eigenvectors  $U_i$  of  $\tilde{\Delta}$  oscillate with frequencies  $\pm\omega_i$ , where  $\omega_i^2$  is the corresponding eigenvalue and the sign corresponds to the phase of  $\dot{U}_i$ .  $\tilde{\Delta}$  and its eigenvector matrix  $U$  are natural candidates for defining wavelets, replacing the Laplace operator and the Fourier transform in conventional wavelet theory. If  $\tilde{V}$  is small compared to the harmonic contribution, this solution is amenable to perturbation theories, and interactions between harmonic modes of strongly differing frequencies will be low. Furthermore, in regular molecular dynamics, the maximum time step for *all* DoFs is limited by the Nyquist sampling rate of the highest frequency. If  $\tilde{V}$  is not small, then another reference needs to be used. In the following, it is assumed that  $\tilde{V}$  is either small or can be properly localized.

## 2.1 Basic Multiresolution

---

Here we introduce the employed wavelet transform and its derivation. We use the multiresolution analysis for diffusion wavelets as introduced by Coifman and Maggioni<sup>6</sup> In essence, the multiresolution decomposition partitions the eigenvalues and eigenvectors of  $\tilde{\Delta}$ , effectively coupling high frequencies in the time domain strongly to high-frequency eigenvectors of  $\tilde{\Delta}$  in the “particle” domain. This is a very important point for the feasibility of the approach because not only can DoFs be reduced, but simultaneously the time step may be increased.

The multiresolution scheme relies on the filter  $T$  and an accuracy operator  $P_\varepsilon$  that projects eigenvectors of a matrix  $X \in \text{span}\{T^{2^n} | n \in \mathbb{N}\}$  with associated eigenvalue less than  $\varepsilon > 0$  to zero. The recursively defined vector spaces

$$V_{n+1} = P_\varepsilon(T^{2^{n+1}})V_n \text{ and} \quad (3)$$

$$W_{n+1} = \ker P_\varepsilon(T^{2^{n+1}}|_{V_n}) \quad (4)$$

are iteratively associated with orthonormal bases via QR-decompositions,

$$T^{2^n} = Q_n R_n,$$

where  $Q_n$  is unitary,  $R_n$  is upper triangular, and both are dependent on the basis used to express  $T^{2^n}$ . At each iteration, the operator is cast in the  $Q$  basis of the previous iteration,

$$\left( \prod_{i=0}^n Q_i \right)^\top T^{2^{n+1}} \left( \prod_{i=0}^n Q_i \right) = Q_{n+1} R_{n+1}.$$

Therefore, repeated application (in infinite precision) is equivalent to the QR algorithm for finding eigenvalues. We separate  $Q_n$  into  $\Phi_n$  and  $\Psi_n$ , where the latter are the columns of  $Q_n$  for which the corresponding rows of  $R_n$  have a norm below  $\varepsilon$ .  $\Phi_n$  collects the remaining columns of  $Q_n$ . Thus,

$$\left( \prod_{i=0}^{n-1} Q_i \right)^\top T^{2^{n+1}} \left( \prod_{i=0}^{n-1} Q_i \right) \approx \Phi_n^\top Q_n R_n R_n^\top Q_n^\top \Phi_n^\top \approx \left( \prod_{i=0}^n \Phi_i^\top \right) T^{2^{n+1}} \left( \prod_{i=0}^n \Phi_i \right).$$

Since  $T$  is positive definite and  $\|T\|_\infty = 1$ , the squaring introduces a *de facto* projection operator  $P_\varepsilon$  via the machine precision. It is this projection that makes the difference between the QR algorithm and the wavelet decomposition.  $W_n$  is spanned by eigenvectors of  $T$  whose eigenvalues obey

$$\varepsilon^{1/2^{n-1}} < \lambda \leq \varepsilon^{1/2^n}. \quad (5)$$

Since the low-pass filter  $T = I - \tilde{\Delta}/C$ , where  $C$  is a sufficiently large constant, the frequency range  $\Delta\lambda = \varepsilon^{1/2^n} - \varepsilon^{1/2^{n-1}} = \varepsilon^{1/2^n}(1 - \varepsilon^{1/2^n})$  for  $\Psi_n$  is drawn ever tighter with each iteration, and  $\lambda - 1 \rightarrow 0$ .  $C = \max_x \|\Delta\|_\infty$  would generally be optimal computationally as  $\Psi_1 \neq \emptyset$ . One advantage of this approach is the inherent permutational invariance of the wavelet spaces and its ability to deal with arbitrary matrices  $T$  with  $\|T\|_\infty \leq 1$ .

$C = \max_x \|\Delta\|_\infty$  would generally be optimal computationally as  $\Psi_1 \neq \emptyset$ . Eigenvalues for graph Laplacians are known to lie in  $[0, 2 \max_i \tilde{\Delta}_{ii}]$ .<sup>7</sup> We have investigated a variety of estimates, but so far the conservative estimate of the largest eigenvalue  $\delta_{max}$  of  $\tilde{\Delta}$  by  $\min\{2 \max_i \tilde{\Delta}_{ii}, \|\tilde{\Delta}\|\}$  has served best.

Assuming that eigenvalues are distributed approximately quadratically,  $\varepsilon = \frac{1}{2}$  would lead to  $\log_2 N$  scales, where  $N$  is the dimensionality of  $\tilde{\Delta}$ . This has not served well as too many DoFs are lumped together due to issues discussed below. Instead, a higher resolution of  $\varepsilon_{\text{machine}}^{\frac{1}{2}}$  is used, and although this wastes some computation on

the first few iterations, it is equivalent to choosing a tolerance on the scale of  $\delta_{max}$ ,  
 $\varepsilon_{\text{effective}} = \varepsilon_{\text{machine}}^{2-m} \approx 1 - \delta_{max}$ .

All MD simulations were performed in LAMMPS.<sup>8</sup> A crystalline model of polyethylene (PE) consisting of 100,800 atoms was simulated with 1-fs time steps and PCFF<sup>9</sup> from 100 K to 500 K in 50-K increments. A 201,936 atom model of amorphous PE was constructed from a cubic supercell in Materials Studio. We used LAMMPS with 3-fs time steps and OPLS<sup>10,11</sup> for alanine dipeptide.

### 3. Results

---

We looked at other avenues to derive wavelets from  $\tilde{\Delta}$ . One choice is to define a shift operator  $\sigma$  in the eigenbasis of  $\tilde{\Delta}$  ordered by eigenvalue, i.e., an eigenvector  $U_i$  with eigenvalue  $\lambda_i$  is mapped to an eigenvector of  $\lambda_{i+1}$ , and a dilation operator  $D$  that projects every other eigenvector to zero. We found no way to derive these without a full computation of the eigenproblem. Similarly, attempts at defining a fast transform via the filter algebra<sup>12,13</sup> generated by  $T$  failed because it required a full solution to the eigenproblem.

Since  $\tilde{\Delta}$  is used to define the wavelets, we discuss its properties in greater detail. These properties have a major impact on the performance of the wavelet transform. Degeneracies in  $\tilde{\Delta}$  reduce the effectiveness of the wavelets based on  $\tilde{\Delta}$ . More information from the potential  $\tilde{V}$  needs to be incorporated to take further advantage. Degeneracies are rampant in large linear homopolymers, but their repetitive structure makes for particularly simple solutions.

$T$  and  $\tilde{\Delta}$  share the same eigenvectors, albeit with reversed order of eigenvalues, so we restrict our discussion to  $\tilde{\Delta}$ .  $\tilde{\Delta}$  is positive semi-definite, and for each set of indices  $J$ ,  $\tilde{\Delta}_{JJ} M^{\frac{1}{2}} \mathbf{1}_J = 0$ , where  $\tilde{\Delta}_{JJ}$  is the square submatrix of  $\tilde{\Delta}$  with indices in  $J$ , and  $\mathbf{1}_J$  is the vector of ones on indices in  $J$  and 0 otherwise. It is possible to block triagonalize  $\tilde{\Delta}$  using transpositions only with diagonal blocks  $\tilde{\Delta}_{J,J}$  and off-diagonal blocks  $\tilde{\Delta}_{J,K}$ , where  $J$  and  $K$  are disjoint and  $J, K, J \cup K$  are contiguous index sets. Without loss of generality, let  $j < k \forall j \in J, k \in K$ .  $\text{rank } \tilde{\Delta}_{JK}$  is generally low because of the low maximum degree of a vertex in  $\tilde{\Delta}$ . If  $\Lambda = \text{span } \tilde{\Delta}_{KJ}$  and  $\tilde{\Delta}_{JJ}$  has a nontrivial, invariant vector space  $\Gamma$  perpendicular to  $\Lambda$ , then  $\Gamma$  is localized to indices preceding  $K$ . Examples include linear homopolymers discussed in detail below, but also disconnected graphs from individual molecules.

Linear homopolymers can be ordered to have block tridiagonal structure where each nonterminal block is a constant  $m \times m$ -matrix for the off-diagonal  $B$  and diagonal  $A$ , respectively,

$$\tilde{\Delta} = \begin{pmatrix} A & B^* & & & \\ B & \ddots & \ddots & & \\ & \ddots & \ddots & B^* & \\ & & B & A & \end{pmatrix}.$$

Furthermore, the off-diagonal block  $B$  consists of a single nonzero entry. The recursive structure implies that  $\tilde{\Delta}$  for a linear homopolymer of  $n$  repeat units can be reordered by a permutation  $\kappa$  of indices such that  $\kappa^T \tilde{\Delta} \kappa = A \otimes I_n + B \otimes \Sigma_n + B^T \otimes \Sigma_n^T$ , where  $I_n$  is the  $n \times n$  identity matrix, and  $\Sigma_n$  is the  $n \times n$ -matrix with all ones on the first subdiagonal only and zeros elsewhere. We assume that the single nonzero entry in  $B$  is  $B_{1,1}$ , connecting only the first entries of the blocks on the diagonal. In this case,  $\tilde{\Delta}$  takes the simple form

$$\begin{pmatrix} C & A_{1,2}I_n & \cdots & A_{1,m}I_n \\ A_{2,1}I_n & A_{2,2}I_n & \cdots & A_{1,m}I_n \\ \vdots & & \ddots & \vdots \\ A_{m,1}I_n & \cdots & \cdots & A_{m,m}I_n \end{pmatrix}.$$

If  $U$  is a unitary matrix of eigenvectors of the tridiagonal, symmetric Toeplitz matrix  $C$  with eigenvalues  $\eta_i = A_{1,1} + 2|B_{1,1}| \cos(i\pi/(m+1))$ , the block diagonal unitary matrix  $I_m \otimes U$  transforms  $\tilde{\Delta}$  into a square block matrix, for which every block is square and diagonal. The eigensystem of  $\tilde{\Delta}$  then comprises eigensolutions to the matrices  $A + (\eta_i - A_{1,1})e_1e_1^T$ . It follows that any eigenvector  $v$  of  $A$  such that  $v_1 = 0$  has an  $n$ -fold degenerate eigenvalue.

$\tilde{\Delta}$  often has many highly degenerate and localized eigenvalues. This fact constitutes a problem since  $\tilde{\Delta}$  provides no further insight into how to subdivide these subspaces within the given wavelet scheme and the external potential  $\tilde{V}$  is not small. Also, numerical solutions are poorly defined as any unitary transform of the subspace is an eigensolution to  $\tilde{\Delta}$ .

**Example.** Any  $CH_2$  group has an associated medium-frequency, highly localized eigenvector of  $\tilde{\Delta}$ .

This follows from the fact that the hydrogens are leafs on the graph, i.e.,

$$\Delta_{CH_2} = \begin{pmatrix} 2K_{CH} & -K_{CH} & -K_{CH} \\ -K_{CH} & K_{CH} & 0 \\ -K_{CH} & 0 & K_{CH} \end{pmatrix}.$$

If the hydrogens are identified with indices  $i$  and  $j$ , then  $\frac{1}{\sqrt{2}}(e_i - e_j)$  is an eigenvector with an associated frequency of a CH vibration.

To address the degeneracy problem, we have considered using alternate relationships. One approach is to compute the conjugacy class of  $T$  and to choose a physically relevant matrix to lift the degeneracy. Some symmetries have known algorithms with decent performance, such as searching for disconnected subgraphs. But in general, it is prohibitively costly to find arbitrary symmetries and then to choose a relevant one, since there are  $n!$  possible permutations of  $n$  indices alone.

Since the QR decomposition is sensitive to the permutations of the matrix, different bases  $\Psi_n$  of  $W_n$  are found depending on preconditioning. This is of particular importance for the high-frequency modes and highly degenerate frequencies. So, another route to approaching the degeneracy issue was to use preconditioning of  $T$  to bias toward localized solutions in the original basis, e.g., balancing distance of nonzeros to the diagonal or backloading high-degree vertices in the QR scheme. This lead to mixed results and was deemed no more reliable than the standard decomposition. Efforts using  $\tilde{V}$  to address the basis problem within  $\Psi_n$  are ongoing.

### 3.1 Adaptive Multiresolution

---

Since one goal of the project is to retain as little information as necessary, yet to reintroduce information when it is needed on the fly, the following discusses the technical and fundamental issues of reconstruction. Reconstruction is the process of adaptively reintroducing detail after dropping information in coarse-graining. We find that reconstruction is systematically possible for numerical as well as analytical coarse-graining hierarchies.

#### 3.1.1 Reconstruction Theory

In order to discuss reconstruction properly, it is first necessary to put coarse-graining into a wider context. In general, a coarsening  $\gamma : \alpha \rightarrow \beta$  is a continuous surjection between 2 topological state spaces  $\alpha$  and  $\beta$  that can be parameterized by

$n > m$  state variables, respectively. If  $\alpha$  is dressed with a probability measure  $P_\alpha : \{X \subset \alpha\} \rightarrow [0, 1]$  to produce a probability space, then  $\gamma$  induces a probability space on  $\beta$  with probability measure  $P_\beta(k) = P_\alpha(\gamma^{-1}(k))$  where  $k \subset \beta$  and  $\gamma^{-1}(k) \subset \alpha$  is the preimage of  $k$ . It is thus possible to select (reconstruct) a precursor for a state  $b \in \beta$  by sampling  $\gamma^{-1}(\{b\})$  with  $P_\alpha$  via the conditional probability  $P(a|b) = P_\alpha(\{a\} \cap \gamma^{-1}(b)) / P_\alpha(\gamma^{-1}(b))$ .

In MD, the state spaces consist of the positions and their associated momenta and thermodynamic state variables, such as temperature or pressure. The probability distributions are Boltzmann distributions that depend on the studied thermodynamical ensemble. In a sequence of coarsenings  $(\gamma_n, \beta_n)$ , it is generally not cost-effective to sample in the largest space  $\alpha$ , and analytical derivations for  $P_{\beta_n}$  are rarely available. In such cases, approximations need to be made. Common solutions in the MD community are probability measures from iterative Boltzmann inversion or (successive) force matching. Hierarchical iteration thereby produces not only probability distributions on the coarser space, but also conditional probability distributions for a fine state mapping to a coarse state. Furthermore, the probability distributions can be used to indicate when a previously undersampled coarse state subspace is encountered, e.g., using an expected improvement measure of the potential  $-\ln P_\beta$  based on the sampled points, for which the trust boundaries can be precomputed.

We now consider hierarchical coarsenings. Let  $\beta_n^\tau : \mathbb{R}^{m_n} \rightarrow \beta_n$  denote a parameterization of  $\beta_n$ . A coarsening is separable if there exist parameterizations  $\beta_n^\tau(t_1^{(n)}, \dots, t_{m_n}^{(n)})$  and  $\beta_{n+1}^\tau(t_1^{(n+1)}, \dots, t_{m_{n+1}}^{(n+1)})$  of  $\beta_n$  and  $\beta_{n+1}$ , respectively, such that there exists a partition of indices  $(t_i^m)_{i=1 \dots a_m} \in \mathbb{R}^{a_m}$ ,  $(t_i^m)_{i=a+1 \dots m} \in \mathbb{R}^{m-a_m}$ , collected as  $A_m$  and  $B_m$ , and continuous surjective mappings  $\mu_{A_n} : A_n \rightarrow A_{n+1}$  and  $\mu_{B_n} : B_n \rightarrow B_{n+1}$  with

$$\beta_n^\tau(\mu_{A_n}^{-1}(t_A), \mu_{B_n}^{-1}(t_B)) = \gamma_n^{-1}(\beta_{n+1}^\tau(t_A, t_B))$$

for  $t_A \in A_{n+1}, t_B \in B_{n+1}$ . We will call a separable coarsening for which neither  $\mu_{A_n}$  nor  $\mu_{B_n}$  are bijections a “fundamental coarsening.” Fundamental coarsenings induce intermediate coarsenings. The state space  $A_n \times B_{n+1}$  is an intermediate state space with the coarsenings  $\gamma : \beta_n^\tau(t_{A_n}, t_{B_n}) \mapsto (t_{A_n}, \mu_{B_n}(t_{B_n})) \in A_n \times B_{n+1}$  and  $\gamma' : (t_{A_n}, t_{B_{n+1}}) \mapsto \beta_{n+1}^\tau(\mu_{A_n}(t_{A_n}), t_{B_{n+1}}) \in \beta_{n+1}$ .

A sequence of separable coarsenings thereby induces a hierarchy of coarsenings

and associated probability measures, induced as described above. In particular, the reconstruction from a fundamental coarsening can be achieved by reconstructing its separable components separately and independently via the conditional probabilities

$$P(t_{A_n} | t_{A_{n+1}}, t_{B_{n+1}}) = P_{\beta_n}((\{t_{A_n}\} \cap \mu_{A_n}^{-1}(t_{A_{n+1}})) \times \mu_{B_n}^{-1}(t_{B_{n+1}})) / P_{\beta_{n+1}}(t_{A_{n+1}}, t_{B_{n+1}}).$$

It is noteworthy that these intermediate probability distributions are available in analytical as well as numerical settings, since a fundamental coarsening has to include proper statistics for the intermediate state space in order to be consistently sampled. Recursive application of conditional probabilities enables concurrent mixed resolutions. Since the construction of modes from the multiresolution analysis produces a hierarchy of frequencies, it induces a hierarchy of coarsenings by dropping successively higher-frequency modes, i.e., by applying the low-pass filter  $P_\varepsilon T^{2^n}$ .

### 3.1.2 A Priori Approximations for Reconstruction

Implementation of reconstruction algorithms as discussed above requires a starting point. In the following paragraphs, methods are proposed for finding good starting points based on  $\tilde{\Delta}$  and other molecular information that is available prior to simulation.

To second order, a quadratic potential around the equilibrium positions of the transformed coordinates approximates the full potential. We assume dominance of harmonic terms, both in the orginal and transformed basis:

$$V_{\text{bond}}(M^{-\frac{1}{2}}U\tilde{r}) \approx \frac{1}{2}(\tilde{r} - \tilde{r}_0)^\top U^\top \tilde{\Delta} U (\tilde{r} - \tilde{r}_0), \quad (6)$$

where  $\tilde{r}$  is the vector of  $\|r_i\|$ . From statistical mechanics the root mean squared deviation from equilibrium of a harmonic oscillator is

$$\sqrt{\frac{kT}{\lambda}},$$

where  $\lambda$  is the force constant. In other words, the higher-frequency components are increasingly found close to their energy minima. This implies that finer scales only have small deviations from their equilibrium positions, while coarser scales may access a much larger space.

We start by approximating  $\|x_i - x_j\|$  by a Taylor expansion around  $r_{ij}^{(0)}$ . This trans-

forms the bond potential  $V_{\text{bond}}$  into:

$$V_{\text{bond}} \approx \frac{1}{8} \sum_{ij} \frac{K_{ij}}{r_{ij}^{(0)2}} \left( \|x_i - x_j\|^2 - r_{ij}^{(0)2} \right)^2.$$

Hence, to find equilibrium distances for  $\tilde{r}$ , we solve the minimization problem

$$\min_{\tilde{r}_0} \sum C_{ij} \left( \|(M^{-\frac{1}{2}} U \tilde{r}_0)_i - (M^{-\frac{1}{2}} U \tilde{r}_0)_j\|^2 - r_{ij}^{(0)2} \right)^2, \quad (7)$$

where  $C_{ij} = \frac{K_{ij}}{r_{ij}^{(0)2}}$ .

**Example.** The 2 nonzero eigenvalues of  $\text{H}_2\text{O}$  correspond to a unique solution for reconstructing  $\text{H}_2\text{O}$ .

The harmonic Laplacian for  $\text{H}_2\text{O}$ ,

$$\tilde{\Delta}_{\text{H}_2\text{O}} = \begin{pmatrix} 2K_{OH}/m_O & -K_{OH}/m_O^{\frac{1}{2}} & -K_{OH}/m_O^{\frac{1}{2}} \\ -K_{OH}/m_O^{\frac{1}{2}} & K_{OH} & 0 \\ -K_{OH}/m_O^{\frac{1}{2}} & 0 & K_{OH} \end{pmatrix},$$

shares the same structure as  $\text{CH}_2$ . The eigenvalues  $\lambda_{0,1,2}$  of this simple matrix are 0,  $K_{OH}$ , and  $(1 + 2/m_O)K_{OH}$ , respectively. Since there are only 2 independent DoFs aside from the center of mass,  $\tilde{r}_i^{(0)}$  is a 2-dimensional system. In 1 dimension, Eq. 7 has 2 solutions for this simple system;  $\tilde{r}_1^{(0)} = \sqrt{2}r_{OH}$ ,  $\tilde{r}_2^{(0)} = 0$ , which leads to a symmetric linear molecule, and  $\tilde{r}_1^{(0)} = 0$ ,  $\tilde{r}_2^{(0)} = r_{OH} \frac{\sqrt{4+2m_O}}{2m_O^{-\frac{1}{2}}+m_O^{\frac{1}{2}}}$ . In 2 dimensions, an angle between these 2 solutions can be assigned. At  $90^\circ$ , the equilibrium structure of water is recovered.

**Example.**  $\text{HCN}$

The harmonic Laplacian for  $\text{HCN}$ ,

$$\tilde{\Delta}_{\text{HCN}} = \begin{pmatrix} \frac{K_{CH}+K_{CN}}{m_C} & -\frac{K_{CH}}{m_C^{\frac{1}{2}}} & -\frac{K_{CN}}{m_C^{\frac{1}{2}} m_N^{\frac{1}{2}}} \\ -\frac{K_{CH}}{m_C^{\frac{1}{2}}} & K_{CH} & 0 \\ -\frac{K_{CN}}{m_C^{\frac{1}{2}} m_N^{\frac{1}{2}}} & 0 & \frac{K_{CN}}{m_N} \end{pmatrix},$$

is no longer as simple as for  $\text{H}_2\text{O}$ , nor are the eigenvalues except 0 simple; for the generalized Amber force field (GAFF), they are 14.7 and 42.8. The numerical

GAFF mode distances in 1 dimension are  $\tilde{r}_1^{(0)} = 3.26$ ,  $\tilde{r}_2^{(0)} = 0.66$  and  $\tilde{r}_1^{(0)} = 2.61$ ,  $\tilde{r}_2^{(0)} = -1.32$ . Knowing that the equilibrium angle between the 2 modes is ca. 80° selects the first solution to reconstruct the equilibrium structure.

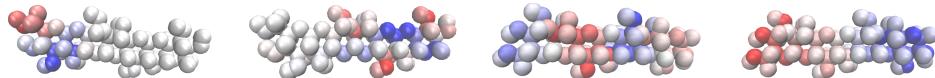
In both examples it was necessary to include angle information to make the best choice. The numerical solution to Eq. 7 can be computed efficiently using a variety of nonlinear least-squares algorithms, but more direct methods are still under investigation. Similar derivations are possible for angle potentials and are under current investigation.

### 3.2 Simulation Results and Analysis

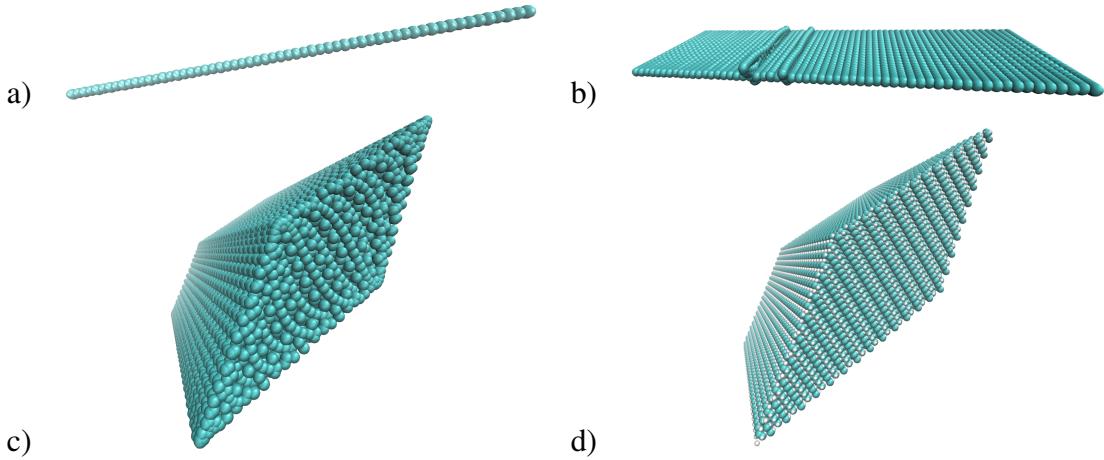
---

We witnessed this behavior in computations of PE as well as polyglucose. Furthermore, when applied to the alanine dipeptide model, commonly used for testing coarse-graining methods for proteins, the wavelet decomposition correctly identified the various chemical groups at the finer scales, while the coarse scales captured the partitioning corresponding to single bond rotations. See Fig. 1 for an illustration on tetra-1,4-D-glucose and Fig. 2 for a demonstration of informational content in the wavelet DoFs for PE.

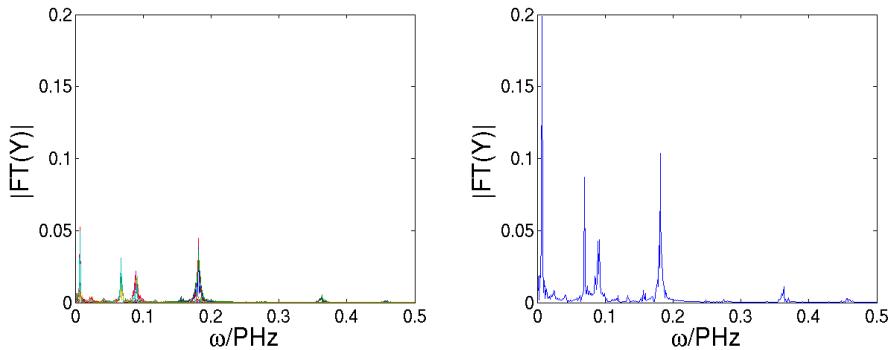
Fig. 3 shows the superimposed Fourier-transformed time series of 1,000 atoms from the 500-K trajectory, which, due to the high temperature, shows the highest mobility of atoms. Although the zero frequency is by far the most intense signal (and is omitted from the figure for clarity), a wide range of other frequencies is active, most importantly around 0.45 PHz, which limits the time step of atomistic simulations of PE to less than 1.2 fs. On the other hand, Fig. 4 shows the effects of scales on time series analysis.



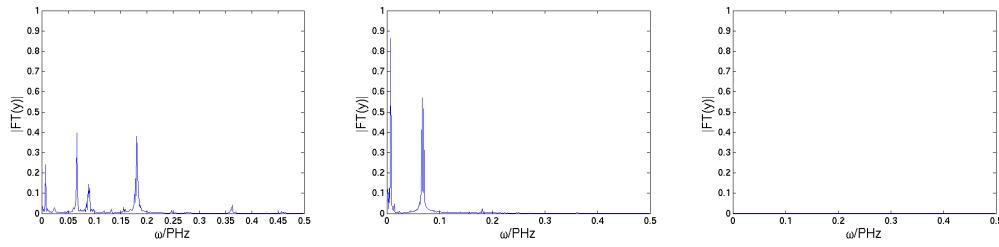
**Fig. 1 Heat map of selected wavelet degrees of freedom (DoFs) in 1,4-D-glucose tetramer. Blue denotes positive coefficients, red negative. From left to right: fine-grained wavelet DoF covering a single repeat unit; coarse-grained wavelet DoF covering 1 half of the oligomer; coarser-wavelet DoF covering the oligomer isolating repeat units with sign changes; coarsest wavelet.**



**Fig. 2** Selected reduced wavelet representations of a polyethylene crystal. From a) to d) wavelet information is successively added. a) Coarse representation highlights the anisotropy in 1-dimensional chain averaged over all chain segments; b) coarse representation of the 2-dimensional chain plane; c) 3-dimensional representation; and d) full resolution.



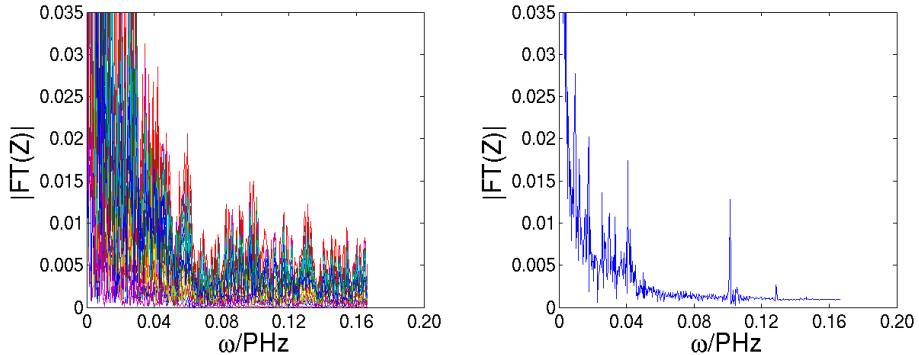
**Fig. 3** Fourier transform of the y-component of 1,000 atoms in crystalline polyethylene (100,800 atoms) excluding zero frequency to allow detail at other frequencies. Molecular dynamics at 500 K and 1 atm. Left: Individual power spectra per atom. Right: Power spectrum of magnitude of optimal representation.



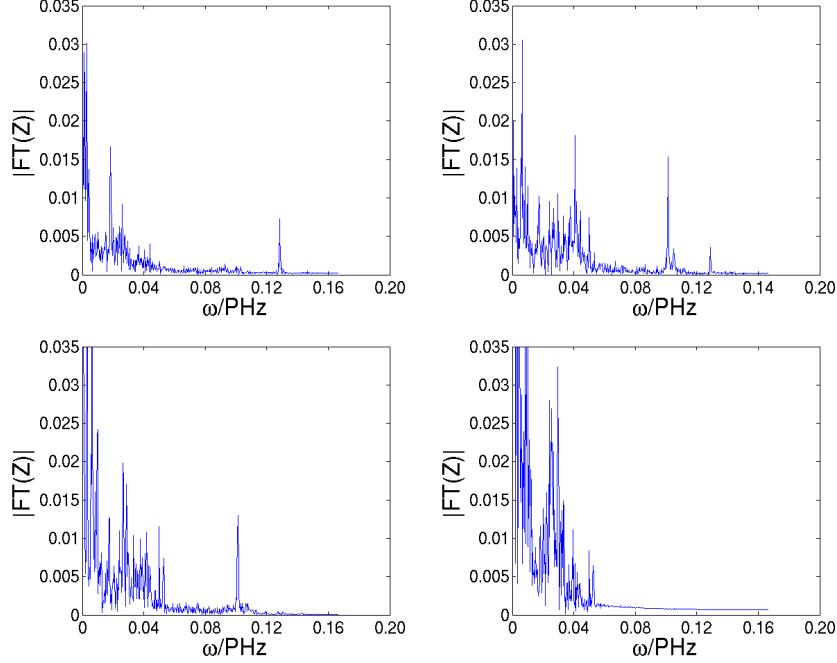
**Fig. 4** Fourier transform of the y-component of a 100,800-atom crystalline polyethylene system sampled at 1 fs. 3 scales are shown: 1st (left, finest scale), 5th (middle), and 12th scale (out of 25, right).

The finest scale out of 25 still retains the high-frequency components (top left), as may be expected, but they are much less intense than the remaining modes. Traversing the scales, we note that a decreasing number of DoFs at the coarser scales shows significant peaks at all. At the medium scale (top right), no high-frequency components are found anymore. Therefore, medium-scale and coarser DoFs are quasi-static compared to the finer scales. Furthermore, the signals are clearly separated and very sharp despite the high temperature, which speaks for strongly decoupled modes and justifies dropping the finer scales, which in turn facilitates speed-up by not only reducing DoFs but also increasing the propagation time step.

The same is witnessed for alanine dipeptide and polyglucose (not shown). Fig. 5 shows the power spectra for all 22 DoFs. Clearly, a large number of frequencies are active for all atoms. The spectrum of optimal representation shows some clear peaks around 0.13 and 0.10 PHz. But in Fig. 6, the 0.13-PHz peak is absent after the 2nd scale, and the 0.10-PHz peak, including noise down to ca. 0.06 PHz, vanishes from the 4th scale on through the remaining 4 scales.



**Fig. 5** Fourier transform of the z-component of alanine dipeptide in vacuum excluding zero frequency to allow detail at other frequencies. Molecular dynamics at 500 K and 1 atm. Left: Individual power spectra per atom. Right: Power spectrum of magnitude of optimal representation.

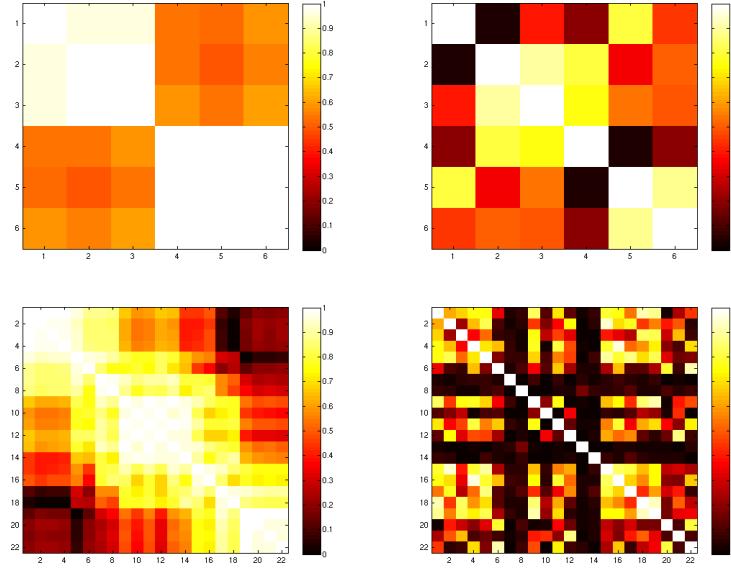


**Fig. 6** Fourier transforms of alanine dipeptide degrees of freedom (DoFs) time series in vacuum excluding zero frequency to allow detail at other frequencies. Molecular dynamics at 500 K and 1 atm. Upper-left: z-component of optimal representation; Upper-right: 2nd finest optimal wavelet DoF z-component; Lower-left: 3rd finest optimal wavelet DoF z-component; Lower-right: 4th finest optimal wavelet DoF z-component.

We investigated the correlation matrices of the various coordinate systems to quantify the extent of interdependence. Fig. 7 shows the correlation matrices,

$$C_{ij} = \frac{\int x_i^*(t)x_j(s-t)dsdt}{\left(\int x_i^*(t)x_i(s-t)dsdt \int x_j^*(t)x_j(s-t)dsdt\right)^{\frac{1}{2}}},$$

where  $x_i(t)$  is a time-dependent coordinate, as heat maps. As may be expected, all particle coordinates are strongly correlated, but in both the water case as well as the alanine dipeptide case, much less correlation is witnessed for the wavelet DoFs (Fig. 7, right), despite the few DoFs involved in these small systems. The extent of overall correlation can be assessed by the  $\ell_1$ -norm of the correlation matrices, ca. 28 versus ca. 21 for water, and ca. 315 versus ca. 168 for alanine dipeptide.



**Fig. 7 Top:** Heat maps of the degree of freedom (DoF) correlations of a water dimer in vacuum; **Bottom:** Heat maps of the DoF correlations of alanine dipeptide in vacuum; **Left:** Correlations of the cartesian DoFs; **Right:** Correlations of the wavelet DoFs.

A further indication of the appropriateness of the diffusion wavelet DoFs is the  $\ell_1$ -norm of structures and forces in a given representation basis. The  $\ell_1$ -norm is bounded by the  $\ell_2$ -norm for any orthonormal basis, and there exists at least one coordinate system in which  $\|x\|_1 = \|x\|_2$ . These coordinate systems optimally describe a state  $x$ .

We compared the  $\ell_1$ -norm in the XYZ coordinate system, the non-lossy wavelet coordinate system, and the optimal coordinate system on each scale. For alanine dipeptide, a reduction by a factor of ca. 2 was achieved by switching to wavelets (3 times the optimum), while the optimal wavelet system managed to reduce to twice the optimum. For polyglucose, the full wavelet DoFs reduce the  $\ell_1$ -norm to half of the XYZ coordinate system, and the optimal wavelet system reduces it to within 20% of the optimum. For crystalline PE, the full wavelet DoFs improve the structural  $\ell_1$ -norm by a factor of ca. 7, while the optimal wavelet representation comes in below 1% of the XYZ coordinate system to a factor of roughly twice the optimal representation. Similar results were obtained for amorphous PE and nanocellulose.

#### 4. Concluding Remarks

---

We have characterized and applied a diffusion-wavelet-based multiresolution of DoFs in MD. The methodology was applied to unimolecular systems of alanine dipeptide, crystalline PE, and an oligomer of 1,4-D-glucose. In all cases, a clear and strong separation of time scales was associated with individual particle scales. This separation was evident even for the very small alanine dipeptide system. Because the particle scales are derived systematically from the underlying bonding topology and do not require any further input, the wavelet decomposition systematically identifies DoFs for conventional coarse-graining procedures. The proposed approach goes beyond conventional approaches based on expert knowledge; however, the wavelet method can be used even in the absence of expert insight, and furthermore, can be used on all length scales, including the smallest. Future work will focus on exploiting these findings to speed up MD a priori.

## 5. References

---

---

1. Lu L, Izvekov S, Das A, Andersen HC, Voth GA. Efficient, regularized, and scalable algorithms for multiscale coarse-graining. *J Chem Theory Comput.* 2010;6(3):954–965.
2. Harmandaris VA, Reith D, Van der Vegt NFA, Kremer K. Comparison between coarse-graining models for polymer systems: Two mapping schemes for polystyrene. *Macromolecular Chem Phys.* 2007;208(2109).
3. Dama JF, Sinit斯基 AV, McCullagh M, Weare J, Roux B, Dinner AR, Voth GA. The theory of ultra-coarse-graining. 1. General principles. *J Chem Theory Comput.* 2013;9(5):2466-2480.
4. Ismail AE, Rutledge GC, Stephanopoulos G. Topological coarse graining of polymer chains using wavelet-accelerated Monte Carlo. I. Freely jointed chains. *J Chem Phys.* 2005;122:234901.
5. Ismail AE, Stephanopoulos G, Rutledge GC. Topological coarse graining of polymer chains using wavelet-accelerated Monte Carlo. II. Self-avoiding chains. *J Chem Phys.* 2005;122:234902.
6. Coifman R, Maggioni M. Diffusion wavelets. *Appl Comput Harm Anal.* 2006;21(1):53–94.
7. Chung FRK. Spectral graph theory. Providence (RI): American Mathematical Society; 1997.
8. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. *J Comput Phys.* 1995;117(1):1–19.
9. Sun H, Mumby SJ, Maple JR, Hagler AT. An ab initio CFF93 all-atom force field for polycarbonates. *J Amer Chem Soc.* 1994;116(7):2978–2987.
10. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Amer Chem Soc.* 1996;118(45):11225-11236.
11. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Chem Phys B.* 2001;105(28):6474-6487.

12. Puschel M, Moura JM. Algebraic signal processing theory: Cooley–Tukey type algorithms for DCTs and DSTs. *Signal Processing, IEEE Transactions on*. 2008;56(4):1502–1521.
13. Voronenko Y, Puschel M. Algebraic signal processing theory: Cooley–Tukey type algorithms for real DFTs. *Signal Processing, IEEE Transactions on*. 2009;57(1):205–222.

## **List of Symbols, Abbreviations, and Acronyms**

---

**CG** coarse-grained

**DoF** degree of freedom

**GAFF** generalized Amber force field

**MD** molecular dynamics

**PE** polyethylene

1 DEFENSE TECHNICAL  
(PDF) INFORMATION CTR  
DTIC OCA

2 DIRECTOR  
(PDF) US ARMY RESEARCH LAB  
RDRL CIO LL  
IMAL HRA MAIL & RECORDS MGMT

1 GOVT PRINTG OFC  
(PDF) A MALHOTRA

1 DIR USARL  
(PDF) RDRL WML B  
B RICE

INTENTIONALLY LEFT BLANK.